DOCUMENT RESUME

ED 275 725 TM 860 639

AUTHOR Lockwood, Robert E.; And Others
TITLE Theoretical Assumptions and Situational Constraints in the Standard-Setting Process.
PUB DATE Apr 86
NOTE 19p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (San Francisco, CA, April 17-20, 1986).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Criterion Referenced Tests; *Cutting Scores; Error of Measurement; High Schools; Minimum Competency Testing; Sampling; *Scores; *Standards; *Statistical Bias; Statistical Distributions; *Test Theory
IDENTIFIERS *Alabama High School Graduation Examination; *Standard Setting

ABSTRACT

Standards, passing scores, or cut scores have been seen as an element of criterion-referenced tests since their introduction. This paper discusses at least two issues surrounding the establishment of cut scores which appear to need clarification: (1) the theoretical definition of a cut score; and (2) decisions which must be made in selecting a procedure within a particular context. Examples from a state testing program show how the theoretical definition of a cut score affects the standards set and why various procedures were selected in specific contexts. (LMO)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED275725

Theoretical Assumptions and Situational Constraints in the Standard-Setting Process

Robert E. Lockwood

Alabama State Department of Education

Glennelle Halpin

Auburn University

James E. McLean

University of Alabama

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☐ This document has been reproduced as received from the person or organization originating it.

☑ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

R. E. Lockwood

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Paper presented at the meeting of the National Council on Measurement in Education, San Francisco, April, 1986

TM 860639

Theoretical Assumptions and Situational Constraints in the Standard-Setting Process

Standards, passing scores, or cut scores have been seen as an element of criterion-referenced tests since their introduction (Glaser, 1963). A general definition of a cut score is that it is the score which divides the persons being tested into those who have mastered that which is being assessed and those who have not.

The issues surrounding the establishment of cut scores or standards have been researched and debated for many years. While a few researchers still oppose the use of standard-setting procedures altogether--a position most prominently championed by Glass (1978)--most researchers have directed their efforts toward defining and refining procedures that may be used to set a cut score or standard (e.g., Angoff, 1971; Ebel, 1979; Huynh, 1976; Jensen, 1978; Nedelsky, 1954; van der Linden & Mellenbergh, 1977). The literature also abounds with comparisons of various procedures and the impacts these particular procedures may have on the standards that result (e.g., Andrew & Hecht, 1976; Brennan & Lockwood, 1980; Halpin & Halpin, 1981; Jaeger, 1979; Koffler, 1980; Livingston & Zieky, 1982; Shepard, 1980).

At least two issues still appear to need clarification. The first of these is the theoretical definition of a cut score. A second area that is becoming critical revolves around the decisions which must be made in selecting a procedure within a particular context. These two issues are addressed in this paper. Examples from a state testing program showing how the theoretical definition of a cut score affects the standards set and why various procedures were selected in specific contexts will be given.

Theoretical Assumptions

There are at least two schools of thought concerning the theoretical definition of a cut score or standard. One assumes that a cut score exists only as it is defined by the procedure used to derive it. Thus, there exists a multiplicity of cut scores, each defined by the procedure used to establish it. With the Angoff approach (Angoff, 1971), for example, the cut score is defined as the probability that a minimally competent examinee will know the correct answer to a particular question, with the sum of the item probabilities being the passing score. In the Nedelsky procedure (Nedelsky, 1954) the cut score is defined as the probability that a minimally competent examinee will guess the correct answer from among options that he/she cannot eliminate as obviously incorrect answers, with the cut score being these chance probabilities summed over all items. Within this school of thought these two procedures define independent cut scores--cut scores that are not estimates of a theoretical parameter but parameters themselves, based on the assumptions of the procedure selected.

A second school of thought assumes that a single "true" cut score exists in theory and that it may be defined as the point on a score scale that separates a master's true score from a nonmaster's true score. Within this school of thought it appears logical to assume that all procedures are attempts to estimate this theoretical value. Unfortunately, it is not possible, at this time, to define the theoretical sampling distribution of estimated values from any or all standard-setting procedures. This difficulty is due, in part, to the fact that what is being estimated has not been precisely defined and consequently the bias that may exist with particular procedures cannot be identified.

If it is assumed that all procedures are estimates of an unknown (and currently unknowable parameter) and that these procedures may or may not contain an unknown degree of bias, what can be specified to help minimize errors of estimation? For any given test, assuming a multiple-choice type test, it is possible to establish reasonable bounds on the true cut score in order to minimize errors.

The bounds on the parameter can be defined as the point below which it can reasonably be assumed that a minimally competent examinee would not score and the point above which it is virtually certain, given measurement error, that the examinee has mastery of what is being tested. The first point, the lower bound, is based on the chance score distribution for the particular test in question and was proposed by Jensen (1978). This point is defined as follows:

$$n_{min} = np + 2.326\sqrt{npq} + \frac{n-(np + 2.326\sqrt{npq}}{k} + 2.326\sqrt{[n-(np + 2.326\sqrt{npq}]pq},$$

where n = number of items on the test,

k = number of options per item,

$p = 1/k$, and

$q = 1-p$.

The resulting score has a 99% probability of being above the distribution of scores that could be obtained by random guessing and indicates some, albeit minimal, amount of knowledge of the domain being tested.

The upper bound is simply perfect performance on the test minus a specified number of items that could be attributed to measurement error and thus is defined as

$$n_{max} = n-2.326\ s_x\sqrt{1-r_{xx}}$$

It is assumed that 99% of those examinees who knew the answer to all questions on the test would have scores above n_{max}.

Because it is assumed that the parameter cut score separates true masters from nonmasters, it appears reasonable to assume that this score would be above n_{min} and below n_{max}. The exact location of the parameter cut score and the shape of the sampling distribution cannot be specified. It is possible, however, to visualize potential sampling distributions from procedures that are negatively biased, positively biased, or unbiased which may be useful in clarifying the issue.

It can be assumed that the sampling distribution of estimates from unbiased procedures would have a mean equal to the parameter. The mean of the sampling distribution of estimates from negatively biased procedures would be less than the parameter and the mean from positively biased procedures would be greater than the parameter. The relationships between these distributions are illustrated in Figure 1. (It will be noted that

---

Insert Figure 1 about here

---

these distributions are shown as mound shaped for discussion purposes; however, it is premature to attempt to specify the shape of these sampling distributions.)

Because the bias that may exist within any procedure cannot be specified, a process should be employed that will minimize the error that would result from using biased procedures. If only one standard-setting procedure is used, the resulting cut score is the best estimate of the parameter. However, if more than one procedure were to be used, the average

of these procedures would minimize the error of estimation, given the constraints of the unknown. This follows because the specific bias of a particular procedure cannot be known and, therefore, it is judicious to assume that both positive and negative biases are represented in the procedures employed. Thus, the mean of the various procedures employed would minimize the error because the direction of bias for each procedure, if any, is not known.

Within the second school of thought, then, there appears to be a rationale for (a) establishing reasonable limits on the cut scores that could be obtained from any procedure used, (b) using as many procedures as practical considering situational constraints, (c) eliminating as unreasonable those obtained cut scores that are outside the range of expected values, and (d) using the mean of the cut scores that were within the expected range.

The theoretical assumption made about the cut score will obviously impact on the choice of which standard-setting procedure is to be used and, subsequently, on the standard that is set. The following example illustrates a standard-setting process that assumed a single true cut score in theory.

Example: Standard Setting on the Alabama High School Graduation Examination

A technical advisory committee composed of persons with measurement expertise from the State Department of Education and state universities used three procedures to set standards on the Alabama High School Graduation Examination (AHSGE) (Technical Advisory Committee, 1984). The first procedure was used to establish the lower bound of potential cut scores for

each subject area of the AHSGE using the approach recommended by Jensen (1978) as previously discussed. These lower bounds were: Reading--49 of 80 items; Language--70 of 119 items; Mathematics--57 of 95 items. Two estimates of the cut score then were obtained using a modified Angoff (1971) procedure and the borderline group method proposed by Zieky and Livingston (1977). For the Angoff procedure the results were: Reading--60 of 80 items; Language--96 of 119 items; Mathematics--62 of 95 items. For the borderline group method the results were: Reading--64 of 80 items; Language--87 of 119 items; Mathematics--61 of 95 items.

With the first school of thought, either of the cut scores derived via the Angoff procedure or the borderline group procedure could be selected as the standard. However, following the assumptions of the second school of thought, the average of the two procedures was selected as a more valid estimate of the true cut score for the AHSGE than either procedure alone. Thus, the resulting standards for the subject area of the AHSGE were: Reading--62 of 80 items; Language--92 of 119 items; Mathematics--62 of 95 items.

Situational Constraints

In many situations where a standard must be established, it is not uncommon for the discussion about procedures to center around which procedure, of many procedures, is best to use. Oftentimes this discussion focuses on the merits of the procedure more than on the situation in which the procedure is to be applied. The choice of a procedure is often dictated by the situation, and there are a number of situational considerations that should be addressed when selecting a procedure.

One of the most crucial considerations is the importance of the decision that will be made about the examinee. Related to the importance of the decision is the consequence of that decision. What is the possible result of classifying an examinee as competent when, in fact, such is not the case or of classifying a competent examinee as not competent. Other concerns are the number of opportunities that an examinee will have to attain a passing score on the test and whether or not the test score will be the sole criterion upon which a decision will be based.

Test characteristics such as type and length of the test as well as the purpose for which the test was designed are also important considerations. Further, examinee characteristics should be taken into account. For example, will a different standard be needed for different subgroups? Resources are another area impacting on the standard-setting process. Are persons available with the expertise necessary to set the standards? Is there sufficient time to do so? Sufficient money?

The following is an example of a standard-setting process wherein the decisions were influenced by situational variables such as these.

Example: Standard Setting by Local Education Agencies on Statewide Basic Competency Tests

Requests from local education agencies were received in the State Department of Education for recommendations regarding pass-fail cut scores for the Basic Competency Tests (BCT) in reading, language, and mathematics that are given statewide to students in Grades 3, 6, and 9. Knowing that these tests had not been designed for pass-fail decisions but rather for the identification of students needing help with specific skills, a technical advisory committee proposed guidelines for establishing multiple cut scores

on the BCT that could be used to identify student mastery status as part of the data to be considered in promotion/retention decisions.

Recognizing the importance of the decision being made plus the fact that examinees would have only one chance to pass, realizing that resources probably would not be available at the local level to set standards, considering the diversity among students across the state, and understanding the nature of the BCT and the competency process, the technical advisory committee recommended that cut scores be established which would place students at Grade 3 and 6 into one of three groups: clearly mastery, questionable mastery, and clearly nonmastery. This decision was, in large measure, dictated by the knowledge that the tests were not designed as pass-fail and thus the BCT alone could only safely classify obvious nonmasters and obvious masters. Based on assumptions and procedures presented earlier, it was felt that the upper and lower bounds of expected cut scores would serve to classify students into the three mastery categories. Students whose scores were below the lower bound could be assumed to be nonmasters. Students whose scores were above the upper bound could be assumed to be masters. For those students whose scores were between these two cut scores, it was felt that the BCT could not classify them without risk of misclassification. Because of the importance of the decision to be made and because students are given only one opportunity to pass the test, this misclassification could not be tolerated. Thus, it was recommended that additional information beyond test performance be used in determining mastery status for students in the questionable category.

Application of the procedures earlier discussed yielded upper and lower bounds for BCT areas in Grades 3 and 6 as shown in Table 1. Based on the

---

Insert Table 1 about here

---

distribution of scores on the BCT in Grades 3 and 6 for the spring, 1985, testing, the percent of students statewide in each of the established categories is given in Table 2. The socio-political climate in the local

---

Insert Table 2 about here

---

districts, no doubt, will be a factor influencing whether standards resulting in percentages such as those in Table 2 will be acceptable.

The same procedure used to establish standards for the BCT at Grades 3 and 6 could have been used for tests at Grade 9. However, since the ninth-grade test is so closely aligned with the Alabama High School Graduation Examination, a different procedure was used. With this procedure--linear regression--scores on the ninth-grade BCT associated with a high probability of passing the graduation examination were derived from the equation

$$\hat{Y} = b_0 + b_1X$$

where $\hat{Y}$ = the predicted AHSGE score

b_0 = the intercept on the y axis

b_1 = the slope of the regression line

X = a score on the ninth-grade BCT.

With this equation and the lower bound of the 95% confidence interval, the score on the BCT associated with first-time success on the AHSGE was identified for each subject area measured by the test. For each subject area these BCT scores were:

Reading	63 of 80 items
Mathematics	68 of 100 items
Language	65 of 77 items.

Having performance on the AHSGE as an established criterion made this standard-setting approach possible. Further, the decision to be made regarding students in Grade 9 is not a pass-fail one but, instead, one of whether or not to place them in a structured remediation program. The question is: Given the probability of passing the AHSGE on his/her first attempt, is the student better served by a more structured remediation program? As was true with the third- and sixth-grade BCT, other information concerning the student should also be considered in answering this question along with other variables such as system personnel, programs, and resources.

Conclusions and Recommendations

The standard-setting processes described together with the discussion of the theoretical assumptions made and situational constraints encountered should prove useful to others who must work their way through the issues and technical matters that have to be considered when selecting and implementing a standard-setting method or methods.

Worthy of reiteration might be the fact that one of the standard-setting processes described herein was recommended in response to requests for official cut scores for basic competency tests which were not designed to be used for promotion/retention. Only when great confidence could be placed in the decision was it suggested that masters/nonmasters be identified based on the BCT alone. Large percentages of students were placed in a questionable category. It was not intended that those students

in the questionable cstegory be considered as students failing to reach mastery. Rather, for these students, the use of data in addition to BCT results was recommended before promotion/retention decisions were made. This approach should result in fewer misclassifications than might be case with decisions based on less information.

For the high school graduation examination, where pass/fail decisions were required, a multimethod approach to standard setting was used yielding diverse standards as has been characteristic of standards set by various methods in other studies. Recommended standards were set by averaging discrepant cut scores based on the assumption that there exists a true cut score separating those who should pass and those who should fail. This average is the best estimate of the true cut score. With this approach, fewer false positives/false negatives are likely to result than might be expected had the respective low/high cut scores been employed. (Both types of errors were considered to be equally undesirable.)

Although it is suggested in this paper that the average cut score produced by several commonly used procedures may be a better estimate of the population cut score than any one individually, much research remains to be done. One of the highest priority research areas should be the derivation of sampling distributions for the more commonly used standard-setting procedures.

References

Andrew, B. J., & Hecht, J. T. (1976). A preliminary investigation of two procedures for setting examination standards. Educational and Psychological Measurement, 36, 35-50.

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), Educational measurement (2nd ed.). Washington, DC: American Council on Education.

Brennan, R. L., & Lockwood, R. E. (1980). A comparison of the Nedelsky and Angoff cutting score procedures using generalizability theory. Applied Psychological Measurement, 4, 219-240.

Ebel, R. L. (1979). Essentials of educational measurement (3rd ed.). Englewood Cliffs, NJ: Prentice-Hall.

Glaser, R. (1963). Instructional technology and the measurement of learning outcomes: Some questions. American Psychologist, 18, 519-521.

Glass, G. V. (1978). Standards and criteria. Journal of Educational Measurement, 15, 237-261.

Halpin, G., & Halpin, G. (1981, November). Reliability and validity of the Ebel, Nedelsky, and Angoff methods of standard setting. Paper presented at Mid-South Educational Research Association meeting, New Orleans, LA.

Huynh, H. (1976). Statistical consideration of mastery scores. Psychometrika, 41, 65-78.

Jaeger, R. M. (1979). Measurement consequences of selected standard-setting models. In M. A. Bunda & J. R. Sanders (Eds.), Practices and problems in competency-based measurement. Washington, DC: National Council on Measurement in Education.

Jensen, O. (1978). Some logical procedures for setting passing points in the absence of direct criterion references. Internal document. Princeton, NJ: Educational Testing Service.

Koffler, S. L. (1980). A comparison of approaches for setting proficiency standards. Journal of Educational Measurement, 17, 167-178.

Livingston, S. A., & Zieky, M. J. (1982). Passing scores: A manual for setting standards of perfo ance on educational and occupational tests. Princeton, NJ: Education Testing Service.

Nedelsky, L. (1954). Abs ute grading standards for objective tests. Educational and Psychol ji al Measurement, 14, 3-19.

Shepard, L. A. (1980). Techr cal issues in minimum competency testing. In D. C. Berliner (Ed.), Revie of research in education (Vol. 8). Washington, DC: American Educational Research Association.

Technical Advisory Committee. (1984, April). A multimethod approach for establishing the passing scores on the Alabama High School Graduation Examination. A symposium presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA. (ERIC Document Reproduction Service No. ED 246 114)

van der Linden, W. J., & Mellenbergh, G. J. (1977). Optimal cutting scores using a linear loss function. Applied Psychological Measurement, 1, 593-599.

Zieky, M. J., & Livingston, S. A. (1977). Manual for setting standards on the Basic Skills Assessment Tests. Princeton, NJ: Educational Testing Service.

Table 1

Score Boundaries Establishing Mastery Status for the BCT

Grade/area	Category/boundary			
	Clearly nonmastery	Questionable mastery		Clearly mastery
Grade 3				
Reading	0	53	80	88
Mathematics	0	38	53	60
Language	0	41	60	67
Grade 6				
Reading	0	48	72	80
Mathematics	0	52	76	86
Language	0	47	70	78

Table 2

Percentages of Students in Each Mastery Category on the BCT

Grade/area	Category		
	Clearly nonmastery	Questionable mastery	Clearly mastery
Grade 3			
Reading	7.9	42.87	49.23
Mathematics	7.6	44.64	47.76
Language	7.3	34.24	58.46
Grade 6			
Reading	7.8	40.87	51.33
Mathematics	19.36	60.63	20.01
Language	7.4	53.46	39.14

Figure 1. Sampling distributions of standard setting procedures.